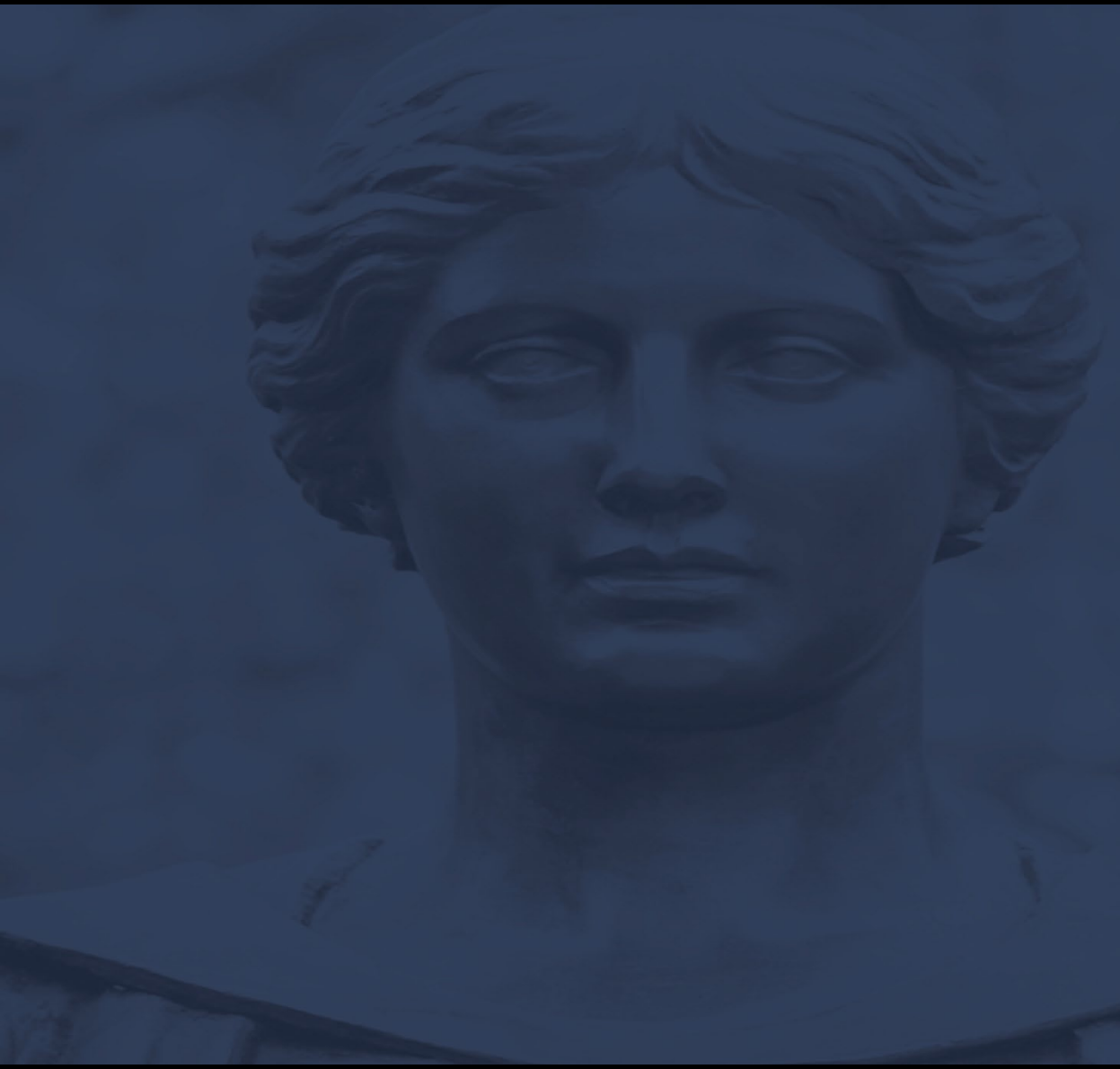
ILLINOIS



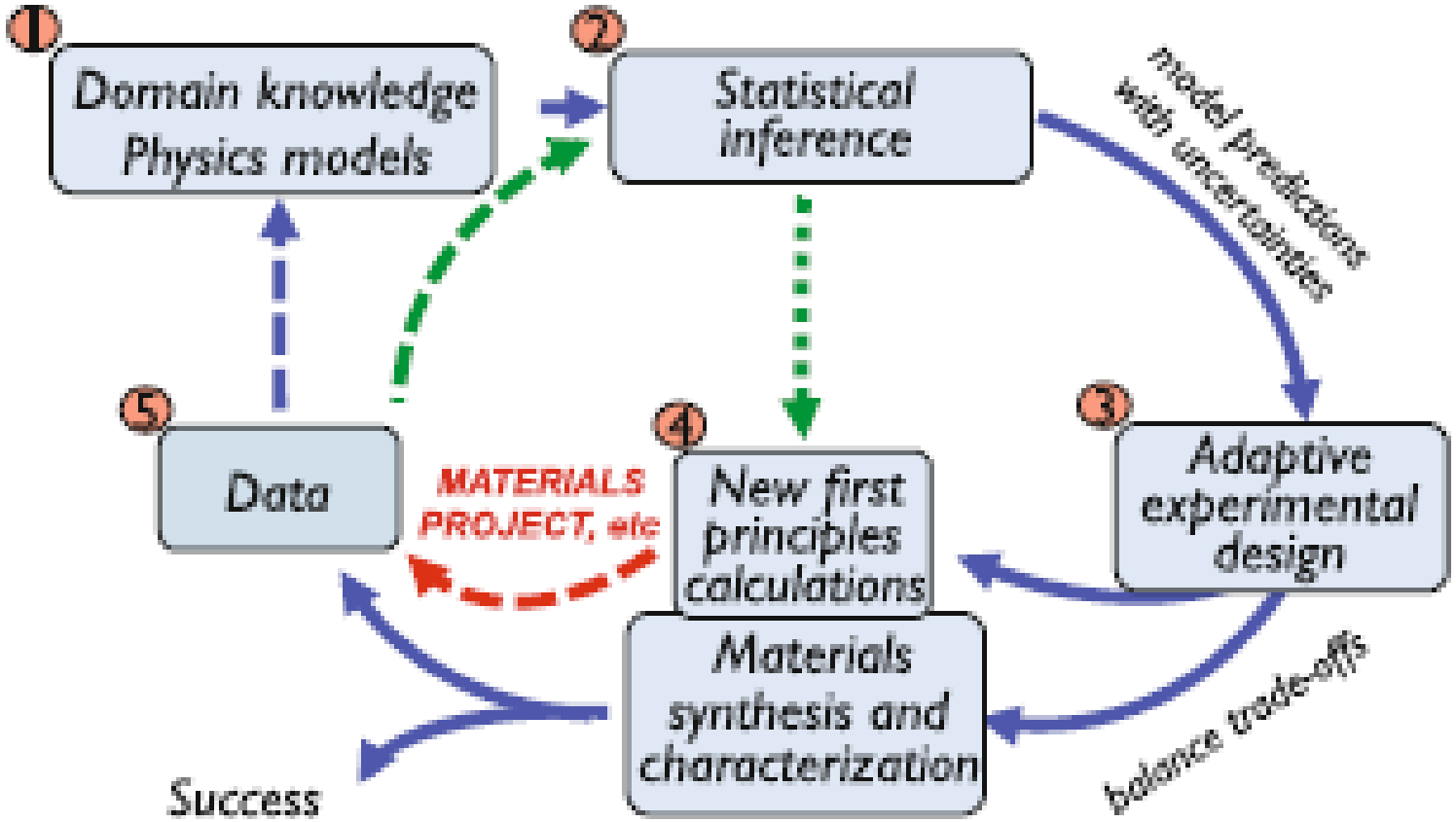
Machine Learning in Material Design

Dr. Yumeng Li

Office: TB 212

Email: Yumengl@illinois.edu

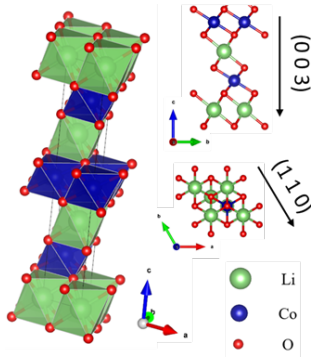
Schematic Illustration of ML in Material Design



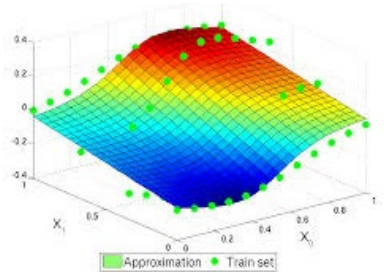
Generic process of building a ML model

Dataset construction

Featurization/ Encoding



Training/Optimization



Evaluation

Data and Databases

- The Five Vs of Data Science
  - Volume: the quantities of data for training and testing
  - Veracity: the quality of data, specifically the degree of accuracy or truthfulness of data sets
  - Validity: related to veracity, the reliability of data sets for a specific application
  - Velocity: the speed at which data flows into a specific database from various sources
  - Volatility: the lifetime of data sets and the rate of change
- Common database types
  - Relational databases: Data are stored in multiple and related tables
  - Document-oriented databases: NoSQLdatabase for unstructured or semistructured data
  - Other databases: cloud databases, key-value databases, and time series databases

Databases

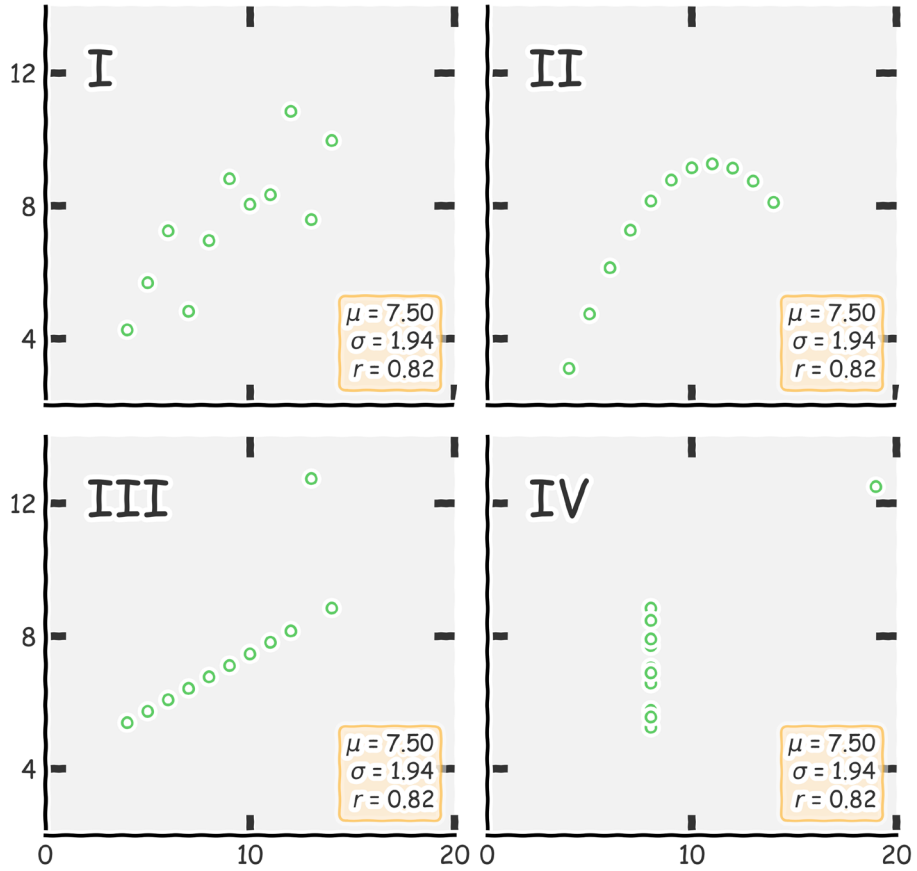
Database	Major material categories	Entries	
Material project	Oxide and Battery materials	711,561	Computational material data
Automatic –Flow for Material Discovery (AFLOW)	Intermetallic materials	3,482,348	
Open Quantum Materials Database (OQMD)		1,022,603	
Material Cloud		7,502,686	
NOvel MAterials Discovery (NOMAD)			
Cambridge Structural Database (CSD)	Small organic molecules and organometallic crystal structures	>1,000,000	Experimental measurements

FAIR

- The FAIR principles (Findable, Accessible, Interoperable, Reusable) and CARE principles (Collective benefit, Authority to control, Responsibility, Ethics) guide responsible data sharing and management.

Data Preprocessing

- Understanding data distributions and relations
  - Diagnostic metrics
    - Average and standard deviation
  - Pair correlations
  - Visualization
- Data preprocessing
  - Normalization and standardization
  - Dimension reduction
  - Algorithm selection
  - Data representation



Representation and representation learning

- Representation here means how the problem is formulated or presented to the computer.

- Supervised learning vs unsupervised learning
  - If data have labels

- Classical machine learning vs deep learning
  - Decision trees, support vector machines, Gaussian process
  - Neural network

Representation

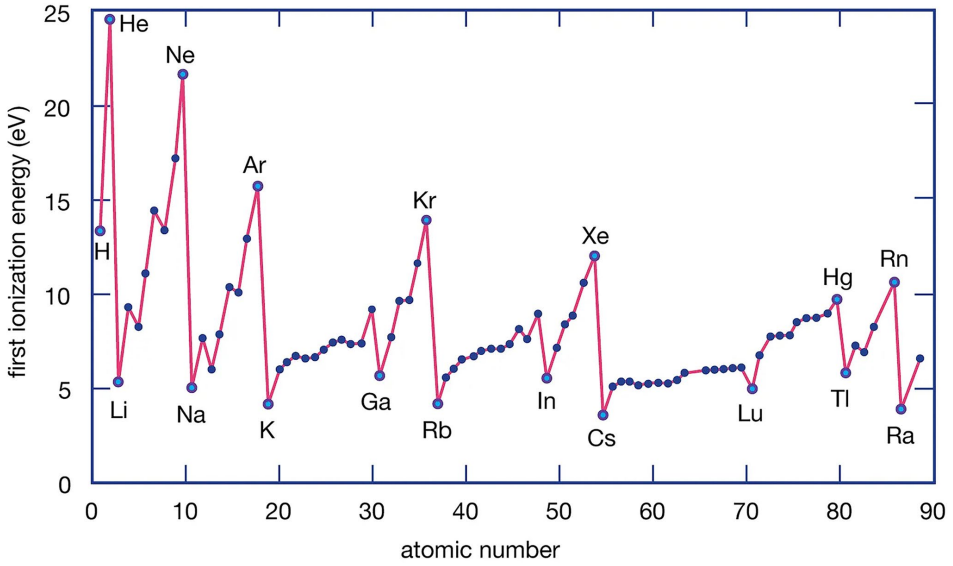


Fig.1 Ionization Energies

Fig.2 Polycrystal microstructure

Evaluation

- Evaluation functions in ML: loss functions or objective functions
  - Loss functions to optimize and calculate the weights and bias

- Evaluation Metrics:
  - MAE:
  - MSE
  - RMSE

  - Cross Entropy
  - Hinge loss
  - Confusion matrix